Tagging for Health Information Organisation and Retrieval

Margaret E. I. Kipp
Faculty of Information and Media Studies
University of Western Ontario
margaret.kipp@gmail.com

NASKO 2007, Toronto, Ontario

Background

- My research examines:
  - how people organise things on the web
  - how this compares to traditional library classification techniques
- Specific points of interest:
  - structures and the creation of structures in classification systems
  - relationship between personal information management and classification

Social Bookmarking and Tagging

- Social Bookmarking:
  - site for sharing bookmarks, articles, etc.
  - association of tags with links
  - tags and articles joined into networks of related terms
  - users encouraged to share links and tags
- Tagging:
  - associating a term with a link or article
  - labelling or classifying for personal use

Tagging and Classification

- mob indexing
- emergent folksonomies
- tag clouds
- grouping by task, by subject, by affective reaction

...

- expert indexing
- planned tree structure of knowledge
- hierarchical
- grouping by subject areas

Previous Studies

- Study 1: Del.icio.us
- study Del.icio.us tag usage on highly tagged sites
- examination of convergence of tag usage
- co-occurrence analysis for co-used tags

- Study 2: CiteULike
- study CiteULike tag usage compared to author keywords and subject headings
- examine types of tags and more traditional index terms

Common Findings

- Study 3: Del.icio.us, CiteULike, Connotea
- use of affective tags (e.g. cool, fun) and time and task related tags (e.g. @toread, todo) in both studies
- > 16% of tags in Del.icio.us study
- average of 1-3 tags per article in original study not directly subject related
- categories: time and task, affective, geographic, methodology, emergent vocabulary, other (no-tag)

Motivations

- Builds on study 2 of CiteULike
- Kipp (2006): users do use words from thesaurus as tags, but often use similar or related terms from other fields
- Examine use of indexing terms by users and indexers
- Do they appear to provide a similar context?

Organisational Structures

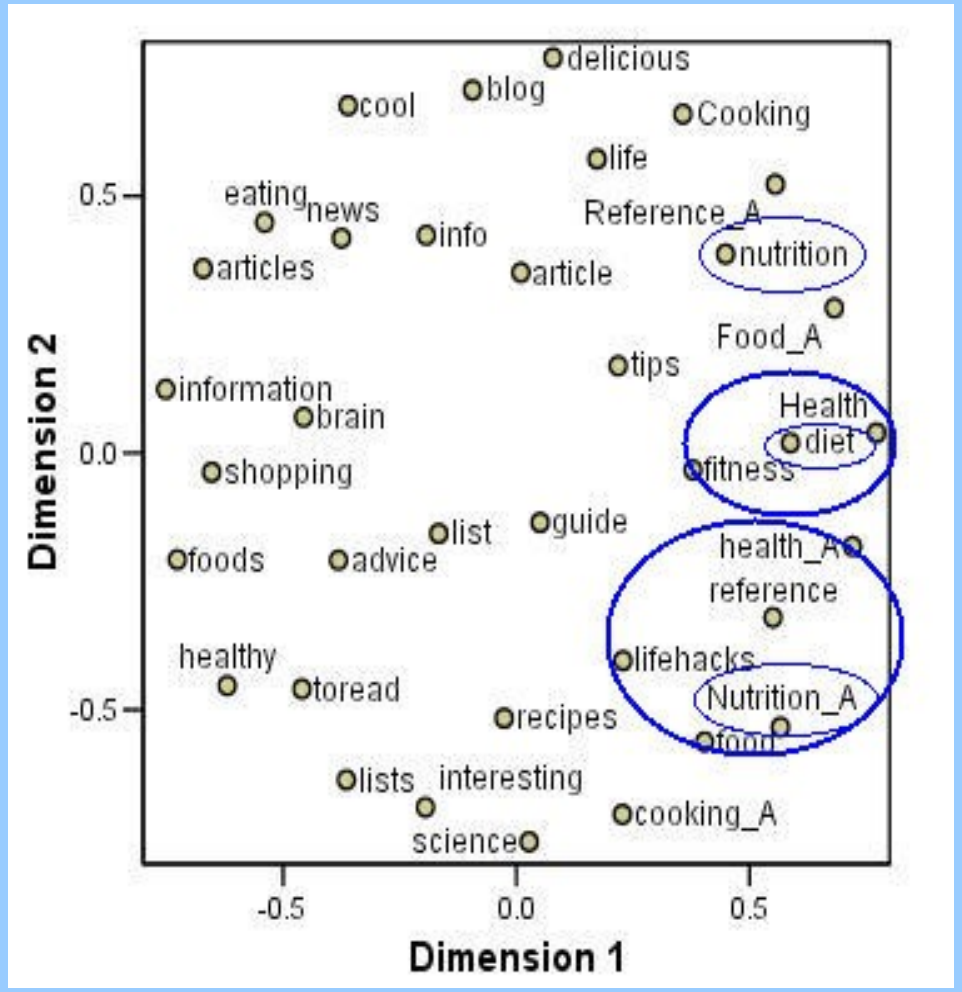
- this study examines the organisational structures emerging in the web 2.0 world
- structures include:
  - tag clouds
  - related tag clusters
  - tag frequency charts
- created structures:
  - co-word graphs of tags

Health Information

- Material:
  - informational pamphlets
  - Health Canada Guides
  - medical journals
  - scientific journals
- Audience:
  - users, patients, families
  - health professionals
  - scientists and researchers in health related fields

Health Information 2

- many user groups; many differing priorities
- some co-word graphs in del.icio.us showed clusters of what might be user groups



Cotag graph www.bellybytes.com

Research Questions

- To what extent do term usage patterns of user tags, author keywords and intermediary descriptors suggest a similar (or differing) context between users and indexers?
- How do tags assigned to health and biology related articles reveal clues to the information context of the taggers?

Data Collection

- three medicine or biology journals:
  - JAMA, Proteins, and J. of Molecular Biology
  - 1 professional journal, 2 academic journals
  - indexed in Pubmed
- 1280 unique articles retrieved from Citeulike
  - 1802 posts (articles may be tagged by multiple users)
- associated Medical Subject Headings (MeSH) collected via Pubmed

Data Analysis

- Informetric analysis using SQL (see Wolfram 2005)
  - standard informetric measures: frequency of occurrence of unique tags
- Thesaural analysis (see Voorbij 1998, Kipp 2006)
  - comparison of terms using Pubmed thesaurus (range from SAME, SYN, NT, BT, RT, related and Not related)

Users and Articles

- Users:
  - 314 unique users, 1802 posts
  - most prolific user had posted 94 posts (median 2)
- Articles:
  - as many as 14 taggers per article (median 3)

User Vocabulary Length

- measure of how many unique tags each user used
- highest number of unique tags used: 18 (min. 1, median 2)
- highest number of unique tags used by a single user: 66 (min. 1, median 4)
- generally connection between high user vocabulary and heavy posting (> 25 articles)

User Vocabulary Length 2

User	Total	Max/Article	Min/Article	Median/Article	Articles Posted
322	66	13	2	8	9
1143	62	8	1	3	94
1005	60	8	1	3	65
3357	54	6	1	3	34
1698	50	9	1	2	34

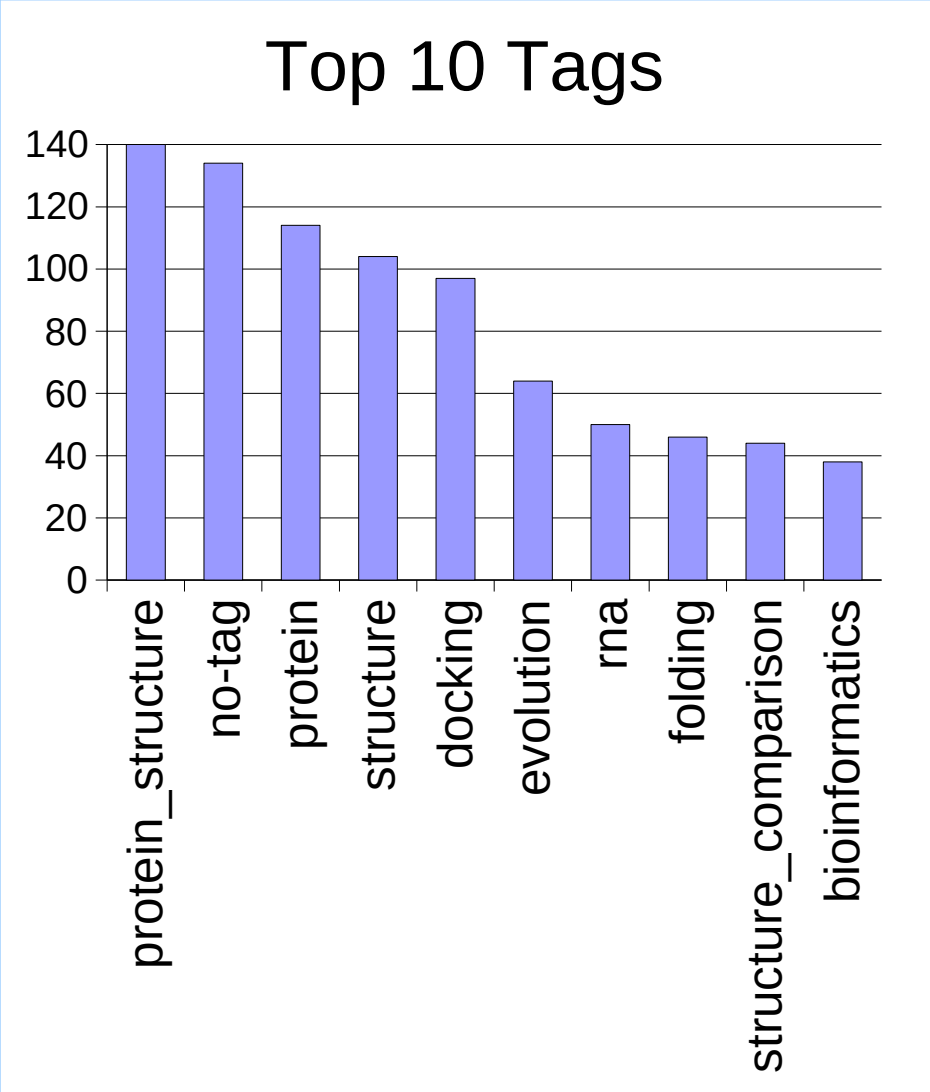
Tags and Descriptors

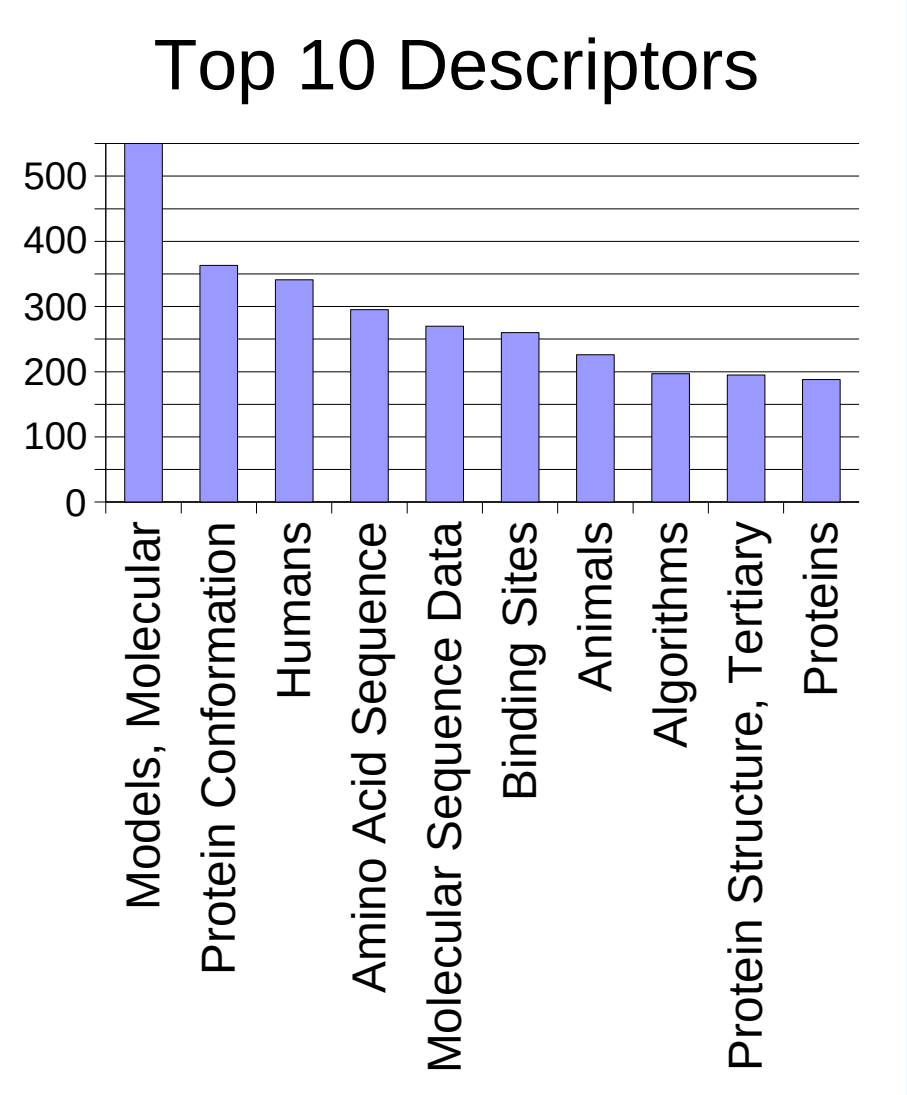
- Tags:
  - 1449 unique tags (total 4289)
  - average 2 tags (max. 29, min. 1)
  - previous studies show users use 1-3 tags
- Descriptors:
  - 2746 unique descriptors (total 14507)
  - average 10 descriptors per article (max. 40, min. 2)

Popular Tags

- popular tags: protein_structure (140), no-tag (134), and protein (114)
- By Journal:
  - docking (Proteins, 85)
  - no-tag (JAMA, 20)
  - protein_structure (J Mol Biol, 52)
- users tagging articles from JAMA do not always assign a tag and may simply be bookmarking their articles

Top 10


Top 10 Tags
140
120
100
80
60
40
20
0
protein_structure
no-tag
protein
structure
docking
evolution
rna
folding
structure_comparison
bioinformatics


Top 10 Descriptors
500
400
300
200
100
0
Models, Molecular
Protein Conformation
Humans
Amino Acid Sequence
Molecular Sequence Data
Binding Sites
Animals
Algorithms
Protein Structure, Tertiary
Proteins

Popular Descriptors

- more heavily reused than tags; tags more likely to be unique
- popular descriptors: 'Models, Molecular', Protein Conformation, and Humans
- By journal:
  - 'Models, Molecular' (Proteins, 252)
  - 'Models, Molecular' (J. Mol. Biol., 385)
  - Humans (JAMA, 137)

Popular Tags by Journal

- no-tag was popular for all journals
- Proteins and Journal of Molecular Biology tags were all related to more basic biological structures:
  - protein_structure, protein, docking, rna
- JAMA tags tended to be more general:
  - cardiology, family-studies, mghlcspub, review

Popular Descriptors by Journal

- Proteins and Journal of Molecular Biology descriptors were related to biological structures:
  - Models, Molecular; Protein Conformation; Amino Acids; Sequence; Proteins
- JAMA descriptors were highly methodology and user group oriented:
  - Humans; Female; Male; Middle Aged

Differences between Journals

- maximum number of keywords (tag or descriptor) per article
  - tags:
    - 29 (Proteins)
    - 20 (JAMA)
    - 19 (Journal of Molecular Biology)
  - descriptors:
    - 40 (JAMA)
    - 36 (Journal of Molecular Biology)
    - 30 (Proteins)

Differences 2

- 6 of 10 articles with highest number of descriptors are JAMA articles
- only 1 of the 10 highest tagged articles is a JAMA article
- users posting JAMA articles tend to use fewer tags, but...
- the more users who post, the higher the number of unique tags per article...

Term Usage

- comparison of tag lists and descriptor lists:
  - many user terms were found to be related to the descriptors but not part of the formal thesaurus
  - may be due to faceting of terms in tags
  - may be due to differing terminology or different view of article emphasis

Title: Optimal diets for prevention of coronary heart disease

- Tags:
  - user1: chd, diet, fat, food, health, heartdisease, lipid, review
  - user2: coronary, diet, disease, heart
- Descriptors:
  - Coronary Arteriosclerosis, Diet, Dietary Carbohydrates, Dietary Fats, Dietary Fiber, Folic Acid, Humans, Life Style, Lipoproteins

Discussion

- results from the previous study (Kipp 2006) using a smaller data set from library science are relevant to other fields and to larger data sets
- users use terminology which is rare or completely absent from descriptor lists (e.g. time and task tags)
- user terms often not part of formal thesaurus

Discussion 2

- Academic versus Professional tagging:
  - distinct difference in tag use between academic journals and professional journal in this study
  - professional tags weighted towards methodology terms, specifically participant groups
  - same phenomenon visible in descriptor usage

Discussions 3

- not everything has to be universal (vertical files, local information)
- user groups may find localised information more useful
- less important to achieve harmony
- more important to achieve access and possible exchange of ideas between user groups

Margaret E. I. Kipp
Faculty of Information and Media Studies
University of Western Ontario
margaret.kipp@gmail.com
http://publish.uwo.ca/~mkipp/

Thank you/Merci!

Questions?